Toward proactive social inclusion powered by machine learning

Emilio Serrano[1] · Mari Carmen Suárez-Figueroa[1] · Jacinto González-Pachón[1] · Asunción Gómez-Pérez[1]

R

Abstract
The fight against social exclusion is at the heart of the Europe 2020 strategy: 120 million people are at risk of suffering this condition in the EU. Risk prediction models are widely used in insurance companies and health services. However, the use of these models to allow an early detection of social exclusion by social workers is not a common practice. This paper describes a data analysis of over 16 K cases with over 60 predictors from the Spanish region of Castilla y León. The use of machine learning paradigms such as logistic regression and random forest makes possible a high precision in predicting chronic social exclusion: around 90% in the most conservative predictions. This prediction models offer a quick rule of thumb that can detect citizens who are in danger of been excluded from the society beyond a temporary situation, allowing social workers to further study these cases.

Keywords Social exclusion · Social services · Data analysis · Machine learning · Data mining

1 Introduction

Social exclusion is a complex and multidimensional process involving the lack of resources, rights, goods and services, and the inability to participate in the normal relationships and activities, available to most people in a society, whether in economic, social, cultural or political scopes [14]. Social exclusion affects not only the quality of life of individuals, but also the equity and cohesion of society as a whole.

✉ Emilio Serrano
emilioserra@fi.upm.es

Mari Carmen Suárez-Figueroa
mcsuarez@fi.upm.es

Jacinto González-Pachón
jgpachon@fi.upm.es

Asunción Gómez-Pérez
asun@fi.upm.es

[1] Ontology Engineering Group, Artificial Intelligence Department, Universidad Politécnica de Madrid, Madrid, Spain

The economic crisis is undermining the sustainability of social protection systems in the EU [6]: 24% of all the EU population (over 120 million people) are at risk of poverty or social exclusion [6]. The fight against poverty and social exclusion is at the heart of the Europe 2020 strategy for smart, sustainable and inclusive growth.

In chronic medical diseases, there is strong evidence supporting that early detection results in less severe outcomes. This paper intends to provide social workers with methods and tools to bring this early detection, which is so beneficial in the medical field, to the challenging problem of chronic social exclusion. Note that although poverty has a significant effect on some dimensions of social exclusion, there are other important causes such as age, ethnicity, disability, gender, and employment status. Therefore, it is considerably more challenging to analyze, detect, treat, and predict social exclusion than poverty.

This paper contributes with an (1) analysis of the social services data of Castilla y León (CyL), which is the largest region in Spain and counts with around two and a half million inhabitants. This analysis allows getting insights into why social exclusion can become chronic. Furthermore, a (2) machine learning model capable of quantifying the risk of chronic social exclusion is build. Finally, a (3) responsive web application is deployed to allow queries by social workers through a number of devices such as smartphones, tablets, or laptops. A RESTful web service is also provided to integrate the predictive capabilities into other software applications.

The paper outline is as follows. After revising some of the most relevant related works in Sect. 2, some of the main methodologies for data mining projects are discussed in Sect. 3. The process followed to analyze the data is explained in Sect. 4. Section 5 reports the outcomes of the experiments conducted. Section 6 explains, analyzes, and compares the results. Section 7 introduces the web service implemented. Finally, Sect. 8 concludes and offers future works. This research work extends a previous conference paper [24].

2 Related works

Prediction models are widely used in insurance companies to allow customers to estimate their policies cost. Manulife Philippines [17] offers a number of online tools to calculate the likelihood of disability, critical illness, or death before the age of 65, based on age, gender, and smoking status. Health is another application field where risk estimations are typical for preventive purposes. More specifically, the risk of heart disease can be estimated at different Web sites such as at the Mayo Clinic Web [18]. The process of gathering and labeling these cases is relatively simple a posteriori. Roughly speaking there is no doubt when someone has suffered one of these conditions.

Some online tools could be used by social services for early detection. Rank and Hirschl [21] give an online calculator that evaluates the probability of experiencing poverty in the next 5, 10, or 15 years based on 4 well-defined fields: race (white or not), education (beyond high school or not), and marital status (married or not). Labeling poverty cases is something automatic when the label or class is defined as falling below a certain annual income.[1] However, the multidimensional nature of conditions such as social exclusion makes considerably more challenging to analyze, detect, treat, and predict it than poverty.

[1] In this vein, the adult dataset [11] is a well-known public labeled dataset that allows predicting whether an adult income exceeds $50 K a year based on a 1994 census database. It can be used to train prediction models as a proof of concept before collecting and labeling the own proprietary data.

A number of data analysis works are important contributions to the use of machine learning in assisting social inclusion. Ramos and Valera [20] use the *logistic regression* (LR) model to study social exclusion in 384 cases labeled by social workers through a manual heuristic procedure. According to this procedure, an individual is considered at a consolidated phase of exclusion if: (1) he or she is living for at least 3 years in unstable accommodation; (2) has very weak links, or none at all, with family or friends; (3) is almost permanently unoccupied; and (4) presents a substantial or total loss of working habits, self-care, or motivation for inclusion. Similar conditions are defined for the initial phase of exclusion. This example of rule of thumb used by the social workers illustrates the complexity and ambiguity of deciding whether someone is suffering social exclusion. Moreover, the heuristic has to be defined before starting gathering data so the social workers can use it. Finally, the authors study a very limited number of cases, less than 400.

Lafuente-Lechuga and Faura-Martínez [12] undertake an analysis of 31 predictors based on segmentation methods and LR. The authors consider the aggregation of scores in different fields related to social exclusion to decide whether a person is under this condition. After a cluster analysis, this score is used to rank and analyze the most important variables to decide whether there is vulnerability to social exclusion.

In a similar style, Haron [9] studies the social exclusion in Israel labeling data by various indicators that are aggregated in a single weighted average score. The author proposes the *linear regression* as a better alternative to the LR. The problem with this approach is that, besides the difficulty in defining these aggregations functions and weights, the machine learning techniques will tend to calculate precisely the aggregation formula since it is defined based exclusively on the training data.

Suh et al. [28] analyze over 35 K cases of 34 European countries using LR. The particular objective of this work is a subjective study and not an objective measure of the social exclusion, for which the researchers use LR over responses to a survey of direct questions about whether people feel excluded from society. Therefore, as the authors point out, there is a subjectivity aspect that is the responsibility of the interviewee instead of the social worker expert.

These inspiring works support the hypothesis that machine learning can greatly benefit social services. Nevertheless, they do not provide social workers with an online tool or an implemented machine learning model to cope with social exclusion. Besides, the number of individuals and the information about each one of these is very limited. Moreover, the use of linear classifiers exclusively such as LR may hinder models from achieving a better predictive power.

3 Methodologies for data mining

Rogalewicz and Sika [23] review methodologies of knowledge discovery and data mining. The main methodologies revised with the high-level phases used to describe the analytics process are the following:

- CRoss-Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM phases are: Business understanding, Data understanding, Data preparation, Modeling, Evaluation, and Deployment.
- Knowledge Discovery in Database (KDD). KDD phases are: Selection, Pre-processing, Transformation, Data mining, and Interpretation/evaluation.

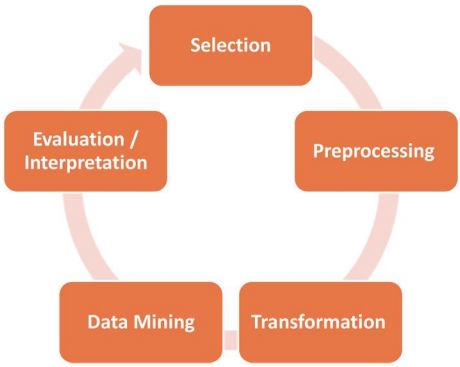

Fig. 1 Knowledge Discovery in Databases methodology

– Sampling, Exploration, Modification, Model, Verification (SEMMA). SEMMA phases are: Sample, Explore, Modify, Model, and Assess.

KDNuggets conducted a poll [19] asking what main methodology voters used for analytics, data mining, or data science projects. The poll included CRISP-DM, KDD, and SEMMA. The votes reflected that CRISP-DM remained the most popular methodology (43% of the 200 votes). However, CRISP-DM is reported to be used by less than 50% voters and there was a significant increase in people using their own methodology (27%). The KDD process was used by 7.5% of the voters. Shafique and Qaiser [27] also revise and compare extensively CRISP-DM, KDD, and SEMMA. The authors conclude that researchers and data mining experts tend to follow the KDD process model, while CRISP-DM and SEMMA are more company oriented. The number of citations to the main references for these methodologies supports this argument: the KDD paper presented by Fayyad et al. [7] counts with over 9900 citations versus less than 900 citations for the CRISP-DM guide [2].

Studying these reviews and although CRISP-DM is an excellent alternative, the KDD process has been chosen for the research presented here. As explained, KDD is one of the three main data mining methodologies in the literature and it is widely used in scientific research. The KDD phases mentioned [7] are displayed in Fig. 1 and described below:

– Step 1: Selection (from data to target data). Selecting data, or focusing on a subset of variables or data samples, to perform discovery on them.
– Step 2: Preprocessing (from target data to processed data). Basic operations include removing noise, collecting the necessary information to model, deciding on strategies for handling missing data fields, and accounting for time-sequence information.
– Step 3: Transformation (from processed data to transformed data). Finding useful features to represent the data depending on the goal of the task. The effective number of variables under consideration can be reduced, or invariant representations for the data can be found.
– Step 4: Data Mining (from transformed data to patterns). Searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering.
– Step 5: Interpretation and/or Evaluation (from patterns to knowledge). Interpreting and evaluating the mined patterns, and possibly returning to any of previous steps for further iteration.

4 Knowledge Discovery in Databases process

As explained, the methodology employed for this data analysis research is the KDD process described by Fayyad et al. [7]. Although the KDD is an iterative and incremental process, some of the decisions made in the different steps are presented unlooped here to make the reading clearer.

4.1 Selection

Eleven databases (DBs) with social services information were available to select relevant data. More specifically, the DBs were implemented with the Oracle object-relational database management system.

After several meetings with the social workers experts, 63 relevant variables from those DBs were selected to further study and preprocess. The predictors were identified by their use in different applications by the social workers. Nonetheless, locating these variables in the DBs to select them was specially challenging because there was not a mapping from the variables to the DB (schema, table, and column). For example, the SAUSS application[2] has a schema with over 800 tables under the hood, plus a large number of tables shared among other applications.

Defining the class to represent a chronic social exclusion situation is another important decision made in this step after several iterations in the methodology. Several prediction services were outlined, but the class was finally defined as "having received social aid during 60 months or more", not necessarily continuously. Intuitively, requiring aid from social services for such a long period involves chronicity of social exclusion. Defining a threshold of 60 months instead of trying to predict the number of months allows the problem to be defined as a binary classification instead of a regression. In our experience, this is an advantage because the evaluation metrics for binary classification are more intuitive for social workers than those used in regression problems, e.g., percentage of correctly classified cases versus Pearson's correlation coefficient. Predicting the economic aid provided to an individual by social services would be another possibility. However, this alternative was soon discarded because a series of small grants may indicate a high degree of social exclusion, e.g., grants for school supplies.

4.2 Preprocessing

The preprocess phase included among others: (1) the data integration where multiple data sources from the selection are combined; (2) the data cleaning removing noise and inconsistent data such as negative income or dates of birth in the year 1900; (3) managing missing values; and (4) generating negative evidence.

Regarding the missing values, a number of variables whose values are missing in over 90% instances were not considered. Moreover, a clear positive correlation between missing data and non-chronic social exclusion was observed in the exploratory data analysis. This supports the idea that these missing values are not random but indicate that the social worker has decided not to log a particular measurement. Therefore, a special value of "NR" (not registered) has been included. As Witten et al. [31] explain, people analyzing medical databases have noticed that cases may be diagnosed simply from the missing values indicating tests

[2] https://sauss.jcyl.es/sauss-sso/.

that a doctor has decided not to make. Imputing values in these cases would result in an information loss.

Concerning the generation of negative evidence, the positive cases are clearly defined as "having received social aid during 60 months or more." However, having received social aid during less than 60 months is not necessarily a negative case because the temporal window of the individual in the social services could make impossible to gather this amount of months. Therefore, there are two extra conditions for negative cases. Firstly, the first registration date for the social patient in the system must be prior to the last 60 months in the databases. Secondly, the last date in which the system logs a follow-up of the social patient must be after the first 60 months recorded in the databases.

As in the main related works studied, see Sect. 2, this proposal is based on *cross-sectional data*, i.e., multiple individuals at the same point of time, or without regard to differences in time. Therefore, there is not a temporal dimension as in *time series*, where a single individual is studied at multiple points in time, or as in *panel data*, where multiple individuals are considered in multiple time periods. Using panel data for a longitudinal study is beyond the scope of this work, but this study would be extremely valuable to evaluate how variables that change over time may affect social exclusion.

4.3 Transformation

In this phase, data are transformed into forms appropriate for mining. This includes, among others: (1) standardization of numeric variables; (2) transforming internal numerical codes into interpretative nominal values; (3) aggregation for the multi-instance learning (where each example in the data comprises several different instances, such as persons with not one but a number of values for a specific variable); and (5) dealing with the imbalanced classification problem.

The result of this process is a dataset with 63 predictors (some of the most relevant ones are described in Sect. 6.3) and 16535 instances: 4205 of the positive class and 12330 of the negative class. This situation is known as imbalanced classification: a high accuracy is achieved by just predicting always the negative class. For example, consider a generic 2-class (binary) classification problem with 100 instances or samples. A total of 20 instances are labeled as "positive" class, and the remaining 80 instances are labeled as "negative" class. This is an imbalanced dataset, and the ratio of positive to negative instances is 20:80, or more concisely 1:4. A simplistic machine learning model could predict new instances with the naive rule "the case is always negative," getting an accuracy of 80%. However, the precision or positive predictive value would be 0% because the model does not predict a single positive case correctly.

There are domains where a class imbalance is not just common but also expected such as the health and medical domains [10,13,30]. For example, consider a machine learning model to predict breast cancer. Except for skin cancers, breast cancer is the most common cancer in American women: one in eight women will suffer this condition throughout their lives [1]. Therefore, the ratio of positive to negative instances expected is 1:8. Note that the problem of the imbalanced classification is not related to the data quality or completeness. In the example, data are supposed to be complete. The imbalance occurs simply because, fortunately, there are more cases of women who do not suffer breast cancer than those who suffer it. In the same manner, social services collect information from many cases and most of them do not present chronicity of social exclusion.

Some approaches to cope with imbalanced classification include: (1) penalized models; (2) undersampling the over represented class (negative); (3) oversampling the underrepresented class (positive); and (4) generating synthetic samples. Section 5 shows several experiments in this vein.

4.4 Data mining

In this phase, machine learning paradigms are applied to create a hypothesis that explains the observations. The *logistic regression* (LR) is widely used to predict the risk of social exclusion as explained in Sect. 2. Furthermore, it is an intuitive solution when a class prediction is wanted with a degree of confidence. Experiments were also conducted using *decision trees* (which typically tolerate imbalanced data) and *rule-based classifiers* (whose hypotheses are highly interpretable for social workers) [25]. Meta-classifiers such as *boosting* and *random forests* (RF) were also considered given their higher predictive power. Besides, these are good out of the box solutions that improve the maintenance when rebuilding new machine learning models in the light of new cases.

4.5 Evaluation

For the evaluation of the models, the cross-validation is typically considered when the performance allows it. Using 10-fold involves rebuilding the machine learning model for the data 11 times. Nonetheless, when oversampling methods are applied, the cross-validation method leads to overoptimistic results since the validation fold may include instances that are also present in the training folds. Thus, the classic partition between training and validation has been undertaken ensuring:

- the splitting (80/20% is considered) preserves the overall class distribution of the data (25% of positive cases);
- and the oversampling is performed after this splitting in both the training and the validation data.

A third partition for testing is not considered given that: (1) the limited positive cases with regard to the negative examples; and (2) the default values for the hyperparameters of the learning algorithms have been used, i.e., validation results have not been employed to optimize hyperparameters.

With regard to the evaluation metrics for the classification, as explained, the *accuracy* $((TP + TN)/(P + N))$ tends to misrepresent the performance when considering imbalanced data. Thereby, *precision* or positive predictive value $(TP/(TP + FP))$ and *recall* or *sensitivity* (TP/P) are more reliable measures.

5 Results

This section describes the results collected after several iterations of the KDD process which is introduced in Sect. 3 and applied in the context of social services in Sect. 4. Table 1 summarizes these results.

Table 1 includes 9 rows for different manners of addressing the imbalanced classification as described in section 4.3. Classification: (1) with the imbalanced data; oversampling the positive class with random sampling with replacement (2) before and (3) after splitting the

Table 1 Experiment results

Experiment	Logistic regression			Random forest		
	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
Imbalanced	81.3	69.3	47.4	80.6	71.3	40.3
Oversampling 1	60.5	78.5	29	91.5	93.1	89.7
Oversampling 2	55.5	78	15.4	67.8	88.6	40.9
SMOTE 1	54.9	82.1	12.5	89.2	96.1	81.8
SMOTE 2	67.9	72.6	57.5	82.4	88.3	74.7
ROSE 1	58.8	56.2	73.4	88	90.7	84.6
ROSE 2	61.4	59	74	73.4	81.7	57.4
Undersampling 1	60.5	62.7	62.7	75.2	77.6	70.9
Undersampling 2	59.9	59	65.3	74.6	78.3	68.1

The model using random forest on Oversampling 2 dataset is referenced as a *conservative model*. The model using random forest on the SMOTE 1 dataset is referenced as an *optimistic model*

Table 2 Comparison of different learning algorithms over Oversampling 2

Algorithm	Accuracy (%)	Precision (%)	Recall (%)
Random forest	67.8	88.6	40.9
Logistic R.	55.5	78	15.4
AdaBoost	68.08	79.9	48.4
C4.5	63.97	75.9	41
RIPPER	72.12	68.1	74.9
k-NN	65.7	73.1	49.8

validation data; oversampling the positive class with SMOTE [3] (4) before and (5) after splitting the validation data; oversampling the positive class with ROSE [16] (6) before and (7) after splitting the validation data; and undersampling the negative class (8) by random sampling with replacement, and by (9) the *K-medoids* segmentation method.

The columns of Table 1 detail experiments for each of the transformed datasets using some of the learning paradigms described in Sect. 4.4. More specifically, a multinomial logistic regression model with a ridge estimator implemented in Weka [31]; and a random forest using the *randomForest* package of R [15] with its defaults parameters, which include the use of 500 decision trees. Finally, different quality metrics for classification introduced in Sect. 4.5 are reported for each learning paradigm: accuracy, precision, and recall.

Table 2 evaluates specific algorithms for different learning paradigms as discussed in Sect. 4.4. Besides the LR and RF already evaluated in Table 1, Table 2 includes: a *decision tree* (C4.5), a *rule-based classifier* (RIPPER), a *boosting* meta-classifier (AdaBoost), and the *k-nearest neighbors* algorithm (with parameter $k = 1$).

6 Discussion

After describing and showing the data collected during experimentation in Sect. 5, this section explains how to interpret these results. For this purpose, this discussion details the baseline, quality metrics, selected machine learning models, and the evaluation. Moreover, the most

relevant variables for predicting the chronic social exclusion and the interpretability of the models are discussed. Finally, the results are compared to the related works, describing the main limitations and benefits of the presented proposal.

6.1 Baseline and quality metrics

The row labeled as *imbalanced* in Table 1 is a baseline for our machine learning models. As explained in section 4.3, there is only one case of chronic social exclusion (positive class) for every three negative cases of this condition. In these experiments, *accuracy* is relatively high: both LR and RF obtain around 80%. However, *precision* and *recall* are lower: LR and RF have a precision of around 70% and a recall of less than 50%.

This situation is known as the *accuracy paradox*, i.e., the accuracy is only reflecting the underlying class distribution. Therefore, accuracy can be a misleading metric for this problem. On the other hand, achieving high precision rates is the principal interest in this research. Considering the chronic social exclusion as the *positive class*, this metric is essential because it allows social workers to find hazardous cases and to focus limited resources on them. Precision is a measure of quality in prediction as recall is a measure of quantity, see Sect. 4.5.

After revising the baseline and the metrics, different approaches for addressing the imbalanced classification are evaluated with the purpose of achieving high precision with an honest validation.

6.2 Selected machine learning models and evaluation

As expected, the oversampling methods offer better results when the validation instances are extracted after oversampling the positive class (rows labeled with *Oversampling 1*, *SMOTE 1*, and *ROSE 1*). In *Oversampling 1*, this happens because several instances that are exactly the same are considered both for training and for validation, see section 4.3. When more advanced methods are employed such as SMOTE and ROSE, new instances are created by generalizating the points where the minority class is valid instead of duplicating cases. Therefore, splitting a validation set after using SMOTE or ROSE is acceptable, although it leads to optimistic results. In these experiments, both LR and RF obtain the best results with *SMOTE 1*: 82.1 and 96.1%, respectively. The random forest with this oversampling method is selected as *optimistic model* and gets the best precision of all the experiments.

Experiments where the validation instances are extracted before oversampling the positive class are labeled with *Oversampling 2*, *SMOTE 2*, and *ROSE 2*. They present a more conservative and honest validation, see Sect. 4.3, because these partitions assure that the same case (or a generalization from several cases) is not in both the training and the validation data. Oversampling the positive class with random sampling, *Oversampling 2*, offers the best results for LR and RF with regard to precision: 78 and 88.6%, respectively. The random forest with this oversampling method is selected as *conservative model*. Figure 2 displays the *Receiver Operating Characteristic* (ROC) curves for the conservative and optimistic models selected.

Undersampling methods, rows labeled with *Undersampling 1* and *Undersampling 2*, not only get worse precision than the oversampling-based alternatives, but also discard a number of negative cases that could reveal interesting characteristics in the study of chronic social exclusion through machine learning.

Regarding the use of different learning paradigms, Table 2 repeats the experiments with the data used for the conservative model (*Oversampling 2*) but changing the learning algorithm.

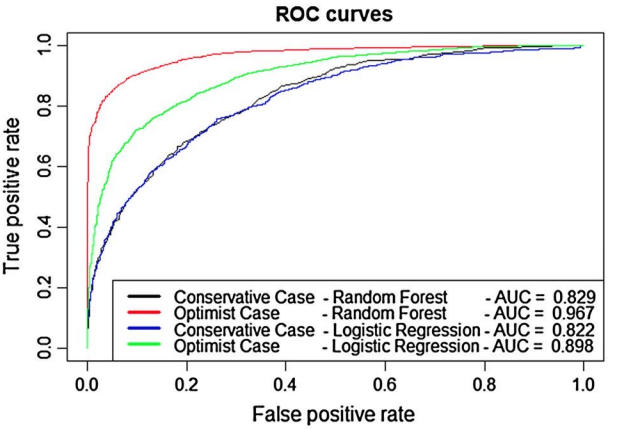

Fig. 2 ROC curves for conservative and optimistic prediction models

As expected, the meta-classifiers random forest and AdaBoost obtain the best results in precision: 88.6 and 79.9%, respectively.

6.3 Feature selection and model interpretation

This section describes a feature selection process over the data, i.e., the process of selecting a subset of relevant features to predict the class. Random forests give variable importance measures to rank variables according to their predictive power in an intuitive manner [4]. Every node in a decision tree is a condition on a single feature that appears higher or lower in the tree according to its relevance for the classification, typically using metrics such as *Gini impurity*, *information gain*, or *entropy*. In addition, since a RF consists of a number of decision trees (500 in our experiments), each feature relevance in each decision tree can be averaged to obtain a raking of the most relevant features. Following this procedure, the ten more important features for predicting chronic social exclusion based on our conservative model are the following:

1. Age: calculated from day of birth to current date or date of death.
2. Level of studies: an ordinal indicator from illiterate to "higher education or vocational training."
3. Classification code: preliminary evaluation label given by the social workers whose values may be temporary, structural, undecided, or (as in most cases) unknown.
4. Annual income in euros: the training dataset contains a great deal of missing data for this variable, but it becomes highly relevant after imputing an average value.
5. Economic activity code: classified by the Spanish Ministry of Employment and Social Security.
6. Civil status.
7. Year of registration in local government.
8. Number of years as job seeker.
9. Professional qualification code: another indicator of education level ordering professional qualifications subject to recognition and accreditation, given by the Spanish Ministry of Education, Culture and Sport.

10. National or foreigner: this binary variable (ternary with the "non-registered" value) was obtained merging a number of nationality codes, most of them too unusual to offer a generic hypothesis about chronic social exclusion.

Note that the validation data have not been used for feature selection, reducing the overfitting risk when using a model trained with only these ten variables. Cross-validation is also desirable in feature selection, but it is overoptimistic when oversampling methods are needed as in this work (see Sect. 4.5). Note also that this feature selection method, which is *embedded* in the RF, selects relevant features without removing redundancies. Some valid alternatives are the use of multivariate filters such as the *Correlation Feature Selection* (CFS) [8] or the use of wrappers.

Regarding the interpretation of the model, age is the most relevant factor for chronic social exclusion and five of the top ten predictors are work or education related. The reader could think of obtaining simple rules as: if the level of studies is one of the most important factors, the higher the level of studies, the lower the risk of chronicity of social exclusion. These are exactly the kind of conclusions obtained by several authors [12,20,28] when using the LR as machine learning model for the analysis and prediction of social exclusion. However, this interpretation is not valid for the RF, not even for a simple decision tree (as C4.5, evaluated in Table 2). Unlike LR, decision trees are not based on defining a constant coefficient for each variable expressing its contribution. In a decision tree, the contribution of each variable depends on the values of other variables that determine the decision path. Furthermore, a RF consists of a large number of deep trees (500 in our experiments), and each tree is trained on bagged data using variables randomly selected. This makes very complex to gain a full understanding of the decision process by examining each individual tree.

However, there is a very good reason to use RF and metaclassifiers even when there is a significant loss of interpretability: they have much more predictive power. As shown in table 1, RF is over 12% more accurate, over 10% more precise, and over 25% more sensitive for the conservative model. For the optimistic model, the increments for the same metrics are around: 34, 14, and 69%. Therefore, our models are much more effective than the LR for generating alerts for the social workers. Moreover, not having an interpretable model does not mean that specific predictions cannot been explained. For a given prediction, the RF can generate the sequence of variables considered for deciding a diagnostic with their weighted contribution, essentially with the same method used here to select relevant variables. We are also considering the inclusion of more complex approaches to understand individual predictions from (powerful) classifiers such as "Local Interpretable Model-Agnostic Explanations" (LIME) [22].

6.4 Comparison, benefits, and limitations

Table 3 shows a comparison of our conservative model with the main related works. As shown, the LR is the most used machine learning model. This learning paradigm is very intuitive to solve a binary classification with a degree of uncertainty. In this way, a LR can predict if a social patient presents: initial or consolidated social exclusion [20]; vulnerability to social exclusion [12]; feeling of social exclusion [28]; or, as in this proposal, chronic social exclusion. Only Haron [9] proposes a linear regression to study if there is correlation between social exclusion and income poverty. One of the advantages of our proposal is the use of RF, a metaclassifier with more predictive power than LR. Besides, the LR and other algorithms have also been used in the experiments presented here, see Table 2.

Table 3 Comparison of the proposal with related works

Research	MLM	Cases	Var.	Acc.	Prec.	Recall	Gen. Ev.	CD
Ramos and Varela [20]	LR	384	5	80.17%	NK	90.37%	NK	NK
Lafuente-Lechuga and Faura-Martínez [12]	LR	NK	31	90.51%	40.61%	NK	NK	NK
Haron [9]	LiR	3600	30	NA	NA	NA	NK	NK
Suh et al. [28]	LR	< 35 K	> 21	NK	NK	NK	NK	NK
CM	RF.	16535	63	67.8%	88.6%	40.9%	Split	4 K/12 K

CM conservative model, *MLM* machine learning model (including LR logistic regression, LiR linear regression, and RF random forest), *Var.* variables, *Acc.* accuracy (including NA not applicable); *Prec.* precision, *Gen. Ev.* generalization evaluation, *CD* class distribution, *NK* not known

Regarding the number of cases analyzed and the number of predictor variables, only Suh et al. [28] suggest the use of more cases or variables than our proposal. However, these are reduced in different population segments without specifying the cases and variables that are finally used to feed the LR. Furthermore, the geographical scope is very different: 34 European countries [28] compared to the Spanish Region of Castilla y León. The sources are also very different: subjective surveys [28] versus anonymized data of social services. Therefore, the second benefit of our work is the quantity and quality of the analyzed data.

The main limitation observed in the reviewed works is that the generalization of the models does not seem to be evaluated, see Sect. 4.5. Thus, the expected result is overfitted models that are not able to predict new cases. Assessing the model accuracy using the same data as in the training phase is always misleading. For example, the k-NN algorithm with $K = 1$ (used in Table 2) would obtain absolute 100% accuracy under these conditions. Besides, these works do not specify the class distribution either. The data analyzed in the social exclusion problem will typically be imbalanced as discussed in Sect. 4.3. Ignoring the imbalanced data leads to an overoptimistic accuracy with a very low precision. Therefore, the third and main benefit of our proposal is the rigorous validation: different datasets are used for building models and validating them, different methods for addressing the imbalanced classification are evaluated, and the precision or positive predictive value is always reported.

The main limitation of our proposal regarding the related works is that the predictive power achieved by RF comes at the cost of losing transparency and interpretability in the model. The LR has a very intuitive interpretation: if the coefficient for the age variable is positive, it means that the risk of chronicity in social exclusion grows with age. This type of "global understanding" of the model is lost by using RFs, boosting, deep learning, or ensembles of these. However, as explained in Sect. 6.3, there are a number of approaches [22,29] to generate human-readable explanations of specific predictions for any classifier. Another disadvantage of our conservative model is that the high precision (88.6%), which is the main goal of our model as explained in section 6.1, comes at the expense of a low recall (40.9%).

7 Implementation

Figure 3 displays the architecture of the current prototype and its integration into the information systems of social services. As shown in the figure, the three-tier client–server software architecture pattern is employed [5]. This allows the three tiers to be upgraded or replaced independently in response to changes in requirements or technology.

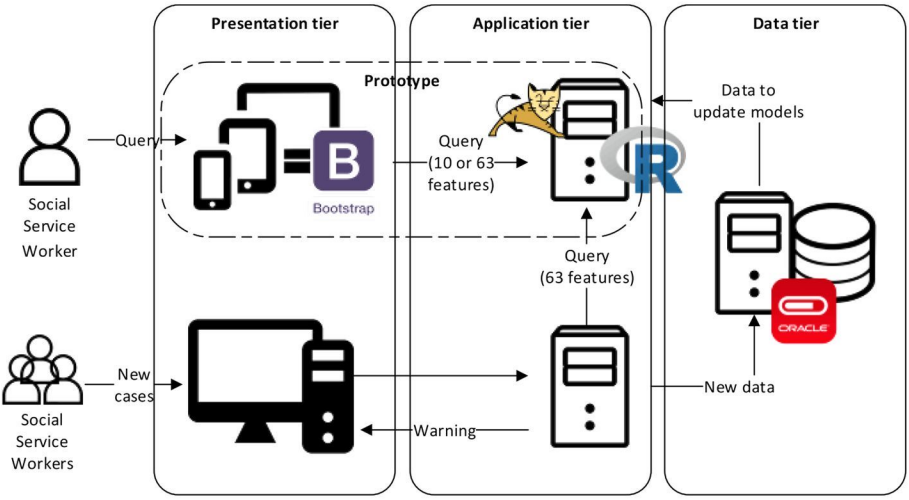

Fig. 3 Architecture of the prototype and its integration into the information systems of social services

Fig. 4 Web application GUI

The *presentation tier* of the prototype is based on a *Bootstrap*[3] web application which is accessible from any web browser. This front-end library ensures the web responsiveness and allows social workers to access the service from a number of devices such as computers, tablets, and smartphones. The interface is shown in Fig. 4. This GUI allows the social worker to type the values of the predictor variables to consult the possible chronicity of a case. Only the ten most important predictors are required by default to improve the GUI usability (see Sect. 6.3). The accuracy, precision, and recall of the conservative model when using only these ten variables are 69.9, 80.9, and 52.2%, respectively. The option of entering all the 63 predictors considered in the machine learning model is also offered. Besides, leaving blank fields that will be taken as unknown values is also allowed. The prediction returns a risk

[3] Bootstrap Web site: https://getbootstrap.com/.

```
{
    "Consulted_case":"55,ESTUDIOS_PRIMARIOS,NR,2195,NR,SOLTERO
        ,1990,6,NINGUNO,S,?",
    "Risk_porcentage":"95.0%",
    "Chronic_case":"Y"
}
```

Fig. 5 JSON response for web service in the application tier

percentage for chronic social exclusion, and it is considered a positive case when the risk is over 50%.

The *application tier* contains the prototype functionality. This tier is based on a Tomcat Server, an open-source Java Servlet Container, that: receives the queries from the presentation tier; consults the machine learning models; and generates a dynamic web content with the predictions. The same queries may be conducted via RESTful web service by introducing the query parameters in the URL. This allows the prototype to provide interoperability between computer systems. Figure 5 shows the JSON output for a RESTful query. The machine learning models are pre-calculated with the R language and stored in this server. The *RCaller* software library allows the R machine learning models to be called from Java.

The complete system proposed in Fig. 3 adds functionality to the application tier, giving it real-time access to social services databases. These databases are currently implemented with *Oracle*. The upgrade would allow the prototype to recover new cases, to automatically preprocess them, and to recalculate the machine learning models. More importantly, the social services applications can use the presented web service when new cases are stored in the databases to consult the models automatically. These applications then can return warnings to social workers if a new case is susceptible to chronic social exclusion.

8 Conclusion and future works

This paper introduces a service to predict the risk of suffering chronic social exclusion with machine learning. With a precision around 90% in the most conservative predictions, it offers a quick rule of thumb that can detect citizens who are in danger of been excluded from the society beyond a temporary situation. The application is available via responsive web and RESTful web service. This allows social workers to consult it from their smartphones and social services software to interact with the application. An early detection is possible thanks to this service, and hence, as in medical diseases, the recovery process can be accelerated.

This service is based on an intelligent model that is fed with data from a whole Spanish region: eleven databases from the social services of Castilla y León (CyL). The classical Knowledge Discovery in Databases (KDD) process has been used and instantiated to the particularities of the data and application field. Some of the main challenges of the analysis are to offer an honest validation and to deal with an imbalanced situation where there is only one case of chronic social exclusion for every four individuals who do not suffer this condition. The results of the analysis reveal the age as the most relevant factor for chronic social exclusion. Besides, five of the top ten predictors are work or education related. Although the web service has been made private temporarily until its integration in the information systems of CyL, both the trained machine learning model and the dataset can be obtained under formal agreement with the Social Services of CyL.

The future works in this research include but are not limited to: extending the prototype with real-time access to social services databases; using panel data for a longitudinal

study; the use of deep learning techniques for feature extraction; the consideration of unlabeled cases to pre-train neural networks before supervised learning; the inclusion of more predictors potentially relevant; the generation of new prediction models by optimizing the hyperparameters of different learning paradigms; offering explanations for the predictions; and addressing security issues as the possibility of extracting personal information from machine learning models [26].

This preliminary research is just the tip of the Iceberg in the potential of artificial intelligence (AI) to assist social services. AI and machine learning can answer a great number of social services-related questions such as: will generational transmission of poverty occurs in this family?; how much economic aid is needed to integrate this person into society?; or how long does it take aid to have an impact on a case?. We aspire to an AI that not only drives our cars or recommends us new series and music, but also provides us with guidelines and suggestions for a greater social inclusion and a happier society.

Acknowledgements This publication would not have been possible without the inputs and collaboration of the Social Services of Castilla y León. This research work is supported by the Spanish Ministry of Economy, Industry and Competitiveness under the R&D project Datos 4.0: Retos y soluciones (TIN2016-78011-C4-4-R, AEI/FEDER, UE).

References

1. American Cancer Society: How Common Is Breast Cancer? (2018). https://goo.gl/5XWwsU (Last Medical Review: July of 2017. Last Revised: January of 2018. Accessed April of 2018)
2. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R (2000) CRISP-DM 1.0 step-by-step data mining guide (tech. rep.). The CRISP-DM consortium. Retrieved from http://www.crisp-dm.org/CRISPWP-0800.pdf
3. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Int Res 16(1):321–357. Retrieved from http://dl.acm.org/citation.cfm?id=1622407.1622416
4. Degenhardt F, Seifert S, Szymczak S (2017) Evaluation of variable selection methods for random forests and omics data sets. Briefings in Bioinformatics, bbx124. Retrieved from https://doi.org/10.1093/bib/bbx124
5. Eckerson WW (1995) Three tier client/server architecture: achieving scalability, performance, and efficiency in client server applications. In: Open Information Systems 10
6. European Commission's DG for Employment, Social Affairs & Inclusion (2018). http://ec.europa.eu/social/main.jsp?catId=751. Accessed April 2018
7. Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) Advances in knowledge discovery and data mining. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds). American Association for Artificial Intelligence, Menlo Park, pp 1–34. Retrieved from http://dl.acm.org/citation.cfm?id=257938.257942
8. Hall MA (1999) Correlation-based feature selection for machine learning (Unpublished doctoral dissertation)
9. Haron N (2013) On social exclusion and income poverty in Israel: findings from the European social survey. In: Berenger V, Bresson F (eds) Poverty and social exclusion around the Mediterranean sea. Springer, Boston, pp 247–269. Retrieved from https://doi.org/10.1007/978-1-4614-5263-8_9
10. Khalilia M, Chakraborty S, Popescu M (2011) Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Dec Mak 11(1):51. Retrieved from https://doi.org/10.1186/1472-6947-11-51
11. Kohavi R, Becker B (1996) Adult data set. https://archive.ics.uci.edu/ml/datasets/Adult. Accessed April 2018
12. Lafuente-Lechuga M, Faura-Martínez U (2013) Análisis de los individuos vulnerables a la exclusión social en españa en 2009. Anales de ASEPUMA(21)
13. Lee J, Wu Y, Kim H (2015) Unbalanced data classification using support vector machines with active learning on scleroderma lung disease patterns. J Appl Stat 42(3):676–689. https://doi.org/10.1080/02664763.2014.978270

14. Levitas R, Pantazis C, Fahmy E, Gordon D, Lloyd E, Patsios D (2007) The multi-dimensional analysis of social exclusion [Book]. Social Exclusion Task Force, Cabinet Office, London. Retrieved from http://www.cabinetoffice.gov.uk/social_exclusion_taskforce/publications/multidimensional.aspx
15. Liaw A, Wiener M (2002) Classification and regression by randomforest. R News 2(3):18–22. Retrieved from http://CRAN.R-project.org/doc/Rnews/
16. Lunardon N, Menardi G, Torelli N (2014) Rose: a package for binary imbalanced learning. R J 6(1):79–89
17. Manulife Philippines. Calculate your risk, your partner's risk or both (2018). http://www.insureright.ca/what-is-your-risk. Accessed April 2018
18. Mayo Clinic. Heart Disease Risk Calculator (2018). http://www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/heart-disease-risk/itt-20084942. Accessed April 2018
19. Piatetsky G (2014) CRISP-DM, still the top methodology for analytics, data mining, or data science projects. https://goo.gl/zrikMB. Accessed April 2018
20. Ramos J, Varela A (2016) Beyond the margins: analyzing social exclusion with a homeless client dataset. Soc Work Soc 14(2):104–120
21. Rank MR, Hirschl TA (2016) Calculate your economic risk. New York times
22. Ribeiro MT, Singh S, Guestrin C (2016) "why should I trust you?" Explaining the predictions of any classifier. CoRR, arXiv:1602.04938. Retrieved from
23. Rogalewicz M, Sika R (2016) Methodologies of knowledge discovery from data and data mining methods in mechanical engineering. Manag Prod Eng Rev 7(4):97–108
24. Serrano E, del Pozo-Jiménez P, Suàrez-Figueroa MC, Gonzàlez-Pachón J, Bajo J, Gómez-Pérez, A (2017) Predicting the risk of suffering chronic social exclusion with machine learning. In: Omatu S, Rodríguez S, Villarrubia G, Faria P, Sitek P, Prieto J (eds) Distributed computing and artificial intelligence, 14th international conference, DCAI 2017, Porto, Portugal, 21–23 June, 2017, vol 620, pp 132–139. Springer. Retrieved from https://doi.org/10.1007/978-3-319-62410-5_16
25. Serrano E, Rovatsos M, Botía JA (2013) Data mining agent conversations: a qualitative approach to multiagent systems analysis. Inf Sci 230:132–146. https://doi.org/10.1016/j.ins.2012.12.019
26. Serrano E, Such JM, Botía JA, García-Fornes A (2014) Strategies for avoiding preference profiling in agent-based e-commerce environments. Appl Intell 40(1):127–142. https://doi.org/10.1007/s10489-013-0448-2
27. Shafique U, Qaiser H (2014) A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). Int J Innov Sci Res 12(1):217–222. Retrieved from http://www.ijisr.issr-journals.org/abstract.php?article=IJISR-14-281-04
28. Suh E, Tiffany Vizard P, Asghar Burchardt T (2013) Quality of life in Europe: social inequalities. 3rd European Quality of Life Survey
29. Turner R (2016) A model explanation system. In: Palmieri FAN, Uncini A, Diamantaras KI, Larsen J (eds) 26th IEEE International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, Vietri sul Mare, Salerno, Italy, pp 1–6
30. Wan X, Liu J, Cheung WK, Tong T (2014) Learning to improve medical decision making from imbalanced data without a priori cost. BMC Med Inform Decis Mak 14:111. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4261533/(111[PII]). DOIurlhttps://doi.org/10.1186/s12911-014-0111-9
31. Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington